

1 **AEM02411-13 REVISED VERSION**

2 **Journal section:** Methods

3 **Title:** GET\_HOMOLOGUES, a versatile software package for scalable and robust  
4 microbial pan-genome analysis

5 **Running Title:** Surveying microbial pan-genomes

6 **Keywords:** bioinformatics, methods, comparative genomics, orthology inference,  
7 *Streptococcus*

8 **Authors:** Bruno Contreras-Moreira<sup>1, 2#</sup> and Pablo Vinuesa<sup>3#</sup>

9 <sup>1</sup>Estación Experimental de Aula Dei (EEAD-CSIC), Avda. Montañana, 1005, 50059  
10 Zaragoza, Spain.

11 <sup>2</sup>Fundación ARAID, calle María de Luna 11, 50018 Zaragoza, Spain.

12 <sup>3</sup>Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México,  
13 Cuernavaca, Morelos, Mexico.

14 <sup>#</sup>To whom correspondence should be addressed.

15

16 Correspondence to:

17 Estación Experimental de Aula Dei-CSIC, Av. Montañana 1005, 50059 Zaragoza, Spain.

18 Phone: +34 976716089. Fax: +34 976716145. E-mail: bcontreras@eead.csic.es

- 19 Centro de Ciencias Genómicas, UNAM, Av. Universidad S/N, Col. Chamilpa, Cuernavaca
- 20 CP 62210, Morelos, México. Phone: +52 777 3175867. Fax: +52 777 3175581. E-mail:
- 21 [vinuesa@ccg.unam.mx](mailto:vinuesa@ccg.unam.mx)

## 22 ABSTRACT

23 GET\_HOMOLOGUES is an open source software package that builds upon popular  
24 orthology-calling approaches making highly customizable and detailed pan-genome  
25 analyses of microorganisms accessible to non-bioinformaticians. It can cluster  
26 homologous gene families using the bidirectional best-hit, COGtriangles or OrthoMCL  
27 clustering algorithms. Clustering stringency can be adjusted by scanning the domain-  
28 composition of proteins using the HMMER3 package, by imposing desired pair-wise  
29 alignment coverage cut-offs or by selecting only syntenic genes. Resulting homologous  
30 gene families can be made even more robust by computing consensus clusters from  
31 those generated by any combination of the clustering algorithms and filtering criteria.  
32 Auxiliary scripts make the construction, interrogation and graphical display of core and  
33 pan-genome sets easy to perform. Exponential and binomial mixture models can be  
34 fitted to the data to estimate theoretical core and pan-genome sizes, and high quality  
35 graphics generated. Furthermore, pan-genome trees can be easily computed and basic  
36 comparative genomics performed to identify lineage-specific genes or gene family  
37 expansions. The software is designed to take advantage of modern multiprocessor  
38 personal computers as well as computer clusters to parallelize time-consuming tasks.  
39 To demonstrate some of these capabilities, we survey a set of 50 *Streptococcus*  
40 genomes annotated in the Orthologous Matrix (OMA) Browser as a benchmark case.  
41 The package can be downloaded at  
42 <http://www.eead.csic.es/compbio/soft/gethoms.php> and  
43 <http://maya.ccg.unam.mx/soft/gethoms.php>.

44

45

46

## 47 INTRODUCTION

48 The ever growing number of sequenced genomes in public databases such as GenBank  
49 has prompted the development of tools aimed at comparing the gene repertoires of  
50 species. Such comparisons include the identification of orthologous genes, assumed to  
51 diverge from a common ancestor after a speciation event, and more likely to conserve  
52 their function across organisms than paralogues (3). For this reason, orthologues are  
53 key elements in genome annotation and evolutionary studies (1, 16). Among bacteria,  
54 which are being sequenced faster than any other domain of life (27), a popular  
55 heuristic recipe for detecting orthologous sequences is simply looking for reciprocal  
56 BLAST hits (4, 28), and different software choices are available for this task (20). By  
57 combining these tools with a growing number of genomic sequences, several recent  
58 studies have provided evidence suggesting that bacterial genomes are actually mosaics  
59 that include genes shared by all isolates of a group of interest (core-genome) and also  
60 strain-specific/partially shared genes (41). The sum of the core genome and the  
61 remaining genes within the group is defined as the pan-genome (40).

62 Here we present GET\_HOMOLOGUES, an open-source software package  
63 released under the GNU General Public license, specifically designed and tested for the  
64 pan-genomic and comparative genomic analysis of bacterial strains at different  
65 phylogenetic distances on Linux/MacOSX computer systems. The software is unique in  
66 several respects. It implements a fully automatic and highly customizable analysis  
67 pipeline, including genome data download, extraction of user-selected sequence  
68 features, running of BLAST and HMMER jobs, as well as indexing, clustering and  
69 parsing of results. It can take advantage of modern multiprocessor architectures, as

70 well as computer clusters, to parallelize time-consuming BLAST and HMMER jobs. It  
71 can handle large datasets (for instance, we have analyzed 101 *Escherichia coli*  
72 genomes) on reasonably modest machines (< 8 GB RAM) by using BerkeleyDB to write  
73 temporary data to disk and/or by calling a heuristic version of our BDBH algorithm.  
74 Auxiliary scripts are integrated to facilitate the parsing and generation of gene families,  
75 including the computation of consensus clusters recovered by combinations of the  
76 sequence clustering algorithms supported. Other scripts are provided for the statistical  
77 analysis and graphical display of results, including core and pan-genome plots, by  
78 calling R functions. Diverse comparative genomics analyses can be also performed.  
79 Finally, an installation script is provided to simplify the installation process and a very  
80 detailed manual with hands-on tutorials is also provided to make this software  
81 package reasonably user-friendly.

82       Here we show some of these capabilities by analyzing a set of 50 *Streptococcus*  
83 genomes downloaded from the most recent version of OMA (Orthologous MAtrix), a  
84 database that identifies orthologs among publicly available, complete genomes (2). We  
85 chose this genus for several reasons. It exhibits very high levels of genome plasticity  
86 (24). The first pan-genomic analyses were conducted on *S. agalactiae* in the pioneering  
87 work of Tettelin and colleagues (39), and very detailed comparative genomics studies  
88 have followed for diverse species in the genus, including the major human pathogens  
89 *S. pyogenes* (23) and *S. pneumoniae* (8), making *Streptococcus* an excellent test case  
90 for the GET\_HOMOLOGUES software.

91

92

## 93 MATERIAL AND METHODS

94 **Input data and output formats.** GET\_HOMOLOGUES takes GenBank or FASTA input  
95 files, and can produce different outputs, as summarized in Figure 1, including  
96 orthologous gene families in FASTA and OrthoXML formats (30), both at the DNA and  
97 amino acid levels.

98 **Third-party software dependencies, data processing and sequence clustering.** The  
99 software is built on top of BLAST+ (6) and the code base of OrthoMCL 1.4 (25), and  
100 supports three popular sequence clustering algorithms: OrthoMCL (OMCL),  
101 COGtriangles (19) and our own implementation of the bidirectional best hit algorithm  
102 (BDBH, see Supplementary Figure S1). Despite their distinct strategies, these  
103 approaches call orthologous sequences using BLAST reciprocal best hits as evidence  
104 (44), and distinguish inparalogues (34) as genes with best hits in the same genome, i.e.  
105 recent paralogues. Moreover, HMMER (<http://hmmer.org>) is integrated to facilitate  
106 Pfam annotation of protein domains (11), so that clusters containing sequences with  
107 different domain architectures, which can confound orthology assignment, can be  
108 filtered out. If input files are in GenBank format, both nucleotide and amino acid  
109 sequence clusters are produced, and orthologous intergenic regions, flanked by  
110 orthologous genes, can also be extracted if required. In addition, genome coordinates  
111 in GenBank files can be used for selecting syntenic clusters, those with at least one  
112 conserved neighbor (Supplementary Figure S2). Finally, as GenBank files contain a  
113 variety of DNA features, the software can also be asked to focus on, for instance, tRNA  
114 genes. In this case BLASTN searches are performed instead of default BLASTP jobs.

115 **Auxiliary scripts for cluster parsing and analysis.** In addition to the main Perl script  
116 *get\_homologues.pl*, this software bundles a few auxiliary scripts to help with  
117 subsequent analyses. For instance, intersection clusters produced by several  
118 algorithms (such as COGtriangles and OMCL) can be easily selected with  
119 *compare\_clusters.pl*, allowing the user to work only with consensus clusters. In  
120 addition, the script *plot\_pancore\_matrix.pl* is provided for plotting pan- and core-  
121 genomes, and fitting the exponential models of Tettelin (40) and Willenbrock (43) to  
122 estimate core and pan-genome sizes. Pan-genomic matrices are conveniently provided  
123 in tabular and PHYLIP format for the automatic generation of pan-genomic trees under  
124 the parsimony criterion, using the PARS program from the PHYLIP package (10), as  
125 illustrated herein. The *parse\_pangenome\_matrix.pl* script is useful for comparative  
126 genomics, focusing on the identification of lineage-specific gene families or  
127 expansions, as well as computing and graphing core, cloud and shell genome  
128 compartments (17). GET\_HOMOLOGUES defines these compartments empirically, as  
129 follows: *Core* - genes contained in all considered genomes/taxa. *Soft core* - genes  
130 contained in 95% of the considered genomes/taxa, as in the work of Kaas and  
131 collaborators (13). *Cloud* - genes present only in a few genomes/taxa. The cutoff was  
132 defined as the most populated non-core cluster class and its immediate neighboring  
133 classes. *Shell*: remaining genes, present in several genomes/taxa. This script will also  
134 compute estimates of the core and pangenomes sizes under the binomial mixture  
135 model of Snipen and colleagues (32). If the source genomic data are provided in  
136 GenBank format, *parse\_pangenome\_matrix.pl* can be asked to plot the clade-specific  
137 genes on a linearized genetic map of a reference genome selected from that lineage.



138 The pan-genomic tree computed by *compare\_clusters.pl* can be useful for selecting the  
139 members of the groups to be compared by *parse\_pangenome\_matrix.pl*.

140 **Benchmark datasets.** In order to test our software pipeline and demonstrate its  
141 capabilities, fifty *Streptococcus* proteomes from 14 species were downloaded in FASTA  
142 format from the Dec2012 release of the OMA Browser (31), together with their  
143 orthologous groups (<http://omabrowser.org>, see Supplementary Table T1). We chose  
144 OMA for this benchmark as it is a validated and updated repository of orthologous  
145 genes across genomes spanning all domains of life (31). OMA is based on an algorithm  
146 that compares genes on the basis of pairwise evolutionary distances instead of BLAST-  
147 scores, considering distance estimation uncertainty and accounts for differential gene  
148 losses (2, 29). In this work we do also re-analyze the 26 *Streptococcus* genomes  
149 analyzed by Lefébure and Stanhope (24) (See Supplementary Table T2).

150 **Proteome annotation.** It was necessary to annotate the OMA clusters, as their  
151 sequence headers contain only OMA identifiers. To do so, we used a modified version  
152 of AutoFact (18), which uses recent, curated and comprehensive sequence databases,  
153 such as NCBI's conserved domain DB (26) and NCBI's Protein Clusters DB (15), in  
154 addition to COG (38), KEGG's PATHWAY DB (14), the Enzyme nomenclature DB  
155 (<http://www.expasy.org/enzyme/>), Pfam (11) and UniRef90 (36) databases.

156 **Supported platforms and availability.** This software is written in Perl and R  
157 (<http://www.R-project.org>) and is best run on a multi-core Linux/MacOSX box or on a  
158 SGE computer cluster (tested with Rocks versions 4.3 and 5.4,  
159 <http://www.rocksclusters.org>). The compressed software package is available for 32  
160 and 64bit processors, includes a user manual with detailed hands-on tutorials, and an

161 installation script which checks for optional software dependencies and guides the  
162 user on how to proceed to install them. This script also supports downloading and  
163 formatting the current version of Pfam. The package can be downloaded at  
164 <http://www.eead.csic.es/compbio/soft/gethoms.php> and  
165 <http://maya.ccg.unam.mx/soft/gethoms.php>.

166

## 167 RESULTS AND DISCUSSION

168 **Comparison of GET\_HOMOLOGUES clusters with those provided by the OMA**  
169 **browser for completely sequenced genomes.** The core-genome for the 50  
170 *Streptococcus* proteomes from 14 species available in the last release of OMA was  
171 calculated with all three clustering strategies (BDBH, OMCL and COG), imposing  
172 minimum pairwise alignment coverage of 75%. Using these parameters, we obtained  
173 487 BDBH, 521 OMCL and 538 COG clusters, and 456 consensus clusters detected by  
174 all three algorithms (see Figure 2A and Supplementary Table T3). The robustness of  
175 GET\_HOMOLOGUES is demonstrated by the fact that these core sets contain all of the  
176 177 core genes reported by the OMA project of orthologous protein families (OMAc),  
177 and many more genes, including 391 BDBH, 413 OMCL and 428 COG clusters with the  
178 same Pfam domain architecture (see Supplementary Table T4). Among these genes  
179 there are essential proteins such as seven 50S ribosomal subunit proteins and six 30S  
180 ribosomal subunit proteins, the translation initiation factor IF-2 protein InfB,  
181 elongation factor G protein Fusa, transcription termination factor protein NusA,  
182 transcription anti-termination protein NusG, DNA topoisomerase protein TopA, ATP-  
183 dependent zinc metalloprotease protein FtsH and recombinase protein RecA, to

184 mention but a few. These are *bona-fide* core genes, most of them present in the  
185 “universal or extended universal” core computed from 12 bacterial and 2 archaeal  
186 phyla (7) (see also Supplementary Table T5). The size of the strict core computed by  
187 GET\_HOMOLOGUES is actually within the range (446-491) of the strict core computed  
188 by Charlebois and Doolittle for 23 *Bacillus/Streptococcus* complete genomes (7),  
189 further highlighting the robustness of this calculation. In addition, sampling  
190 experiments similar to those carried out by Tettelin *et al.* (39), in which *Streptococcus*  
191 genomes are randomly added to the pan-genome pool in order to track the fraction of  
192 unique and common genes contributed, were performed with BDBH to estimate the  
193 theoretical core genome size (Figure 2B). The fitted functions converged to values of  
194 517 and 595 when using the Willenbrock (43) and Tettelin (39) fits, respectively, clearly  
195 much larger than OMAc. As to the pan-genome (Figure 2C), it seems to converge to a  
196 linear growth, as already observed by Tettelin for 8 *S.agalactiae* genomes (39). Notice  
197 however that, as expected, the slope of the pan-genome curve fitted to the larger and  
198 taxonomically more diverse OMA dataset is almost twice (slope = 60.5) that reported  
199 by Tettelin (slope = 33). The pan genomes obtained by OMCL and COG agree for 5398  
200 clusters, but disagree mainly for clusters of less than three genomes, as COGtriangles  
201 does not resolve them (Figure 2D). Panels 2E and 2F show the partition of OMCL pan-  
202 genome clusters. Note that for 50 *Streptococcus* strains, the soft-core compartment  
203 includes sequences found in at least 47 genomes.

204

205 **Computing soft core-genomes with GET\_HOMOLOGUES on datasets containing draft-**  
206 **genomes.** The performance of GET\_HOMOLOGUES core-genomes was further

207 validated with 26 genomes, from 6 *Streptococcus* species, previously analyzed by  
208 Lefébure and Stanhope (24), which (currently) contains 6 draft-genomes. Of the 611  
209 core genes reported by these authors, the consensus OMCL & COGtriangles soft-core  
210 genome recovered 529, as illustrated in Supplementary Figure S3. Soft-core genes are  
211 by definition required to be present in at least 95% of the input genomes, allowing for  
212 missing or fragmented genes, which are expected when comparing datasets that  
213 include draft-genomes (13). A strict core genome for this dataset, based on the  
214 consensus between the BDBH, COG and OMCL algorithms and default 75% pairwise  
215 alignment coverage, contains only 412 gene families. This figure is clearly an  
216 underestimation of the core-genome size, as judged from the estimations obtained  
217 with 50 proteomes from fully sequenced genomes presented in Figure 2B and  
218 Supplementary Table T4. This misleading result is caused by missing or incomplete  
219 genes in the 6 draft genomes included in the dataset, demonstrating the value of  
220 defining a “soft” core-genome (13) and the flexibility of the GET\_HOMOLOGUES  
221 package to adapt to different types of datasets and analysis requirements.

222

223 **Use of pangenome trees to aid in group selection for comparative genomics.** The  
224 auxiliary script *compare\_clusters.pl* can be asked to generate a pan-genomic matrix of  
225 presence-absence of genes, which is inferred from the clusters generated by the main  
226 script *get\_homologues.pl* when run with the option -t 0 (which retrieves clusters of all  
227 sizes, as opposed to default core clusters). Such a matrix was calculated for the 50  
228 *Streptococcus* proteomes extracted from OMA, and then used to calculate the  
229 parsimony pan-genomic tree shown in Figure 3. This kind of trees show the

230 phylogenetic relationship of proteomes based on their gene-family contents. Such  
231 phylogenies arguably reflect better the genetic affinities of the proteomes based on  
232 their composition (presence-absence of homologous genes) and therefore, their  
233 phenotypic potential, than conventional species phylogenies estimated from  
234 concatenated alignments of core genome genes or gene products (33). For comparison  
235 purposes, we also computed a maximum likelihood tree out of 364 concatenated  
236 alignments of the corresponding “strict” single-copy consensus core genes  
237 (Supplementary Figure S4), reported by all BDBH, COGtriangles and OMCL algorithms  
238 with the Pfam domain-scanning option enabled (-D, the corresponding core genome  
239 estimates are shown in Supplementary Table T4). In both phylogenies the species  
240 appear as monophyletic entities. However, when comparing the relationships between  
241 species in both phylogenies, some differences are found. For example, in the pan-  
242 genome tree *S. equi* strains appear as the sister group to *S. pyogenes*, while on the  
243 core-genome phylogenetic hypothesis *S. dysgalactiae* would be the closest relative to  
244 *S. pyogenes*, as also found by Lefebure *et al.* (23) in a similar analysis. Yet in another  
245 study (37), the *S. dysgalactiae* strains analyzed therein grouped as the sister clade of *S.*  
246 *pyogenes* when using hierarchical clustering (UPGMA) based on dissimilarities in gene  
247 content. GET\_HOMOLOGUES uses the parsimony optimality criterion to compute pan-  
248 genomic trees because the accuracy of the UPGMA clustering algorithm is well known  
249 to degrade with increasing deviation of the underlying distance data from  
250 ultrametricity (9). Given the important differences in the rates of gene gain or loss  
251 frequently observed between closely related bacterial lineages (23), deviations from  
252 ultrametricity should be frequently observed in pan-genomic matrices. However, the  
253 tab-delimited pan-genome matrix can be used to compute distance matrices in R or

254 other packages. The differences between the groupings revealed by both types of core  
255 and pan-genome trees can be exploited as alternative evolutionary hypotheses to  
256 guide the selection of species or proteome groups for comparative genomics of clade  
257 pairs.

258 **Identifying clade-specific genes and gene-family expansions.** The auxiliary script  
259 *parse\_pangenome\_matrix.pl* can be used to report cloud, shell, soft core and core  
260 clusters (17) and display them graphically (if R is installed), as shown in Figures 2E and  
261 2F, which indicate that the shell-genome component is the largest one (3232 genes  
262 present in 4-46 strains) for the 50 streptococci analyzed, followed by the cloud  
263 genome (2619 genes present in  $\leq 3$  genomes). These data clearly illustrate the complex  
264 structure of the *Streptococcus* pan-genome. The same script is also useful for  
265 performing basic comparative genomics studies, specifically to identify lineage-specific  
266 genes or gene expansions in a target group “A” with regard to a second reference  
267 group “B”. Option -P can be used to define the percentage of genomes that must  
268 comply with the presence/absence of a particular cluster (the default is 100%) in order  
269 to define genes specifically found or expanded in the focal lineage “A”. As an example,  
270 we performed an analysis to identify such gene families among the 11 *S. pyogenes*  
271 proteomes from the OMA50 dataset, using as a close reference group the *S. equi*  
272 strains. Group selection was based on the pan-genomic tree shown in Figure 3. The  
273 analysis was first performed using the default option -P 100 to identify those gene  
274 clusters present or amplified in all 11 *S. pyogenes* strains. We found 42 *S. pyogenes*-  
275 specific gene clusters and 3 that were selectively expanded in this lineage in  
276 comparison to *S. equi*. Among the first class of genes were well established

streptococcal virulence factors like pyrogenic exotoxins, which are associated with streptococcal toxic shock syndrome and scarlet fever (22), and an hyaluronic acid hydrolase, involved in the degradation of connective tissue and pathogen spread (35). Other annotated proteins include salivaricin A precursor, a lantibiotic produced by >90% oral *S. pyogenes* strains found in human saliva (42), a putative ATP-binding cassette (ABC) transporter previously shown to be involved in virulence of *S. pneumoniae* in a mouse model of infection (5). All these genes were also found as part of the *S. pyogenes* gene repertoire in a recent study by Lefébure and colleagues (23). An additional 13 *S. pyogenes*-specific gene clusters were found when -P was relaxed to 90 (i.e. a cluster was considered lineage-specific when at least 90% of the genomes in the target group A contained it). Among them were bacteriocin UviB, a plasmid stabilization system toxin protein, and the two-component sensor kinase DpiB. The possibility of relaxing this value is particularly useful when dealing with draft genomes, because it makes this type of comparative genomics analysis robust even when missing genes are expected due to incomplete genome sequencing. Supplementary Table T6 shows the complete list of the genes found in these analyses along with their annotations.

294

**Software design and performance benchmarks.** Unlike similar tools (12, 21, 45), which require the user to provide pre-computed sequence similarity results, GET\_HOMOLOGUES explicitly takes care of BLAST and optional Pfam searches. The software can take advantage of modern multiprocessor architectures to parallelize expensive BLAST+ and hmmscan analyses, or submit jobs to a computer cluster (see

300 manual and Supplementary figure S5). For instance, the pan-genome of 20 *Escherichia*  
301 *coli* genomes (93,612 genes) can be analyzed in <6 hours on a commodity laptop (4 GB  
302 RAM, 4 cores). Supplementary Figure S6 plots the memory footprint and runtime of  
303 increasingly larger sequence sets, clearly indicating that the software is more scalable  
304 when using BerkeleyDB, a high-performance embedded database. For instance, 101  
305 *Escherichia coli* genomes can be successfully processed with modest RAM  
306 requirements invoking -s option in the command line: 880MB, 932MB and 902MB for  
307 BDBH, OMCL and COGtriangles, respectively. For extremely large datasets,  
308 GET\_HOMOLOGUES also includes a heuristic option to minimize the number of BLAST  
309 searches required for a core-genome BDBH job, which grows linearly instead of  
310 quadratically (see Supplementary Figure S7 and Table T7).

311 In conclusion, we have shown that GET\_HOMOLOGUES is a powerful, highly  
312 customizable and fully automatic analysis pipeline that makes robust and rigorous pan-  
313 genomic and comparative genomic analyses much easier to perform by microbiologists  
314 without strong bioinformatics skills or dedicated hardware. In addition,  
315 GET\_HOMOLOGUES is scalable, being able to deal with dozens of genomes on  
316 relatively modest computer systems and will handle hundreds of genomes on more  
317 powerful servers or computer clusters, making it suitable for large-scale pan-genomics  
318 and comparative genomics studies. It is open source software freely available for  
319 academic use.



## 320 **ACKNOWLEDGEMENTS**

321 We thank Romualdo Zayas and Víctor del Moral at CCG-UNAM for their technical  
322 support. We also thank Dr. David M. Kristensen and the development team of  
323 OrthoMCL for their permission to use their code in our project. Funding: Fundación  
324 ARAID, Consejo Superior de Investigaciones Científicas [grant number 200720I038].  
325 DGAPA-PAPIIT UNAM-México [grant number IN212010], CONACyT-México [grant  
326 numbers P1-60071 and 179133].

## 327 REFERENCES

- 328 1. **Altenhoff, A. M., and C. Dessimoz.** 2012. Inferring orthology and paralogy. *Methods*  
329 *Mol. Biol.* **855**:259-279.
- 330 2. **Altenhoff, A. M., A. Schneider, G. H. Gonnet, and C. Dessimoz.** 2011. OMA 2011:  
331 orthology inference among 1000 complete genomes. *Nucleic Acids Res.* **39**:D289-294.
- 332 3. **Altenhoff, A. M., R. A. Studer, M. Robinson-Rechavi, and C. Dessimoz.** 2012. Resolving  
333 the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in  
334 function than paralogs. *PLoS Comput. Biol.* **8**:e1002514.
- 335 4. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J.**  
336 **Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database  
337 search programs. *Nucleic Acids Res.* **25**:
- 338 **Basavanna, S., S. Khandavilli, J. Yuste, J. M. Cohen, A. H. Hosie, A. J. Webb, G. H. Thomas,**  
339 **and J. S. Brown.** 2009. Screening of *Streptococcus pneumoniae* ABC transporter  
340 mutants demonstrates that LivJHMGF, a branched-chain amino acid ABC transporter,  
341 is necessary for disease pathogenesis. *Infect. Immun.* **77**:3412-3423.
- 342 6. **Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L.**  
343 **3389-402.**
- 344 5. **Madden.** 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421.
- 345 7. **Charlebois, R. L., and W. F. Doolittle.** 2004. Computing prokaryotic gene ubiquity:  
346 rescuing the core from extinction. *Genome Res.* **14**:2469-2477.
- 347 8. **Donati, C., N. L. Hiller, H. Tettelin, A. Muzzi, N. J. Croucher, S. V. Angiuoli, M. Oggioni,**  
348 **J. C. Dunning Hotopp, F. Z. Hu, D. R. Riley, A. Covacci, T. J. Mitchell, S. D. Bentley, M.**  
349 **Kilian, G. D. Ehrlich, R. Rappuoli, E. R. Moxon, and V. Masignani.** 2010. Structure and  
350 dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species.  
351 *Genome Biol.* **11**:R107.
- 352 9. **Felsenstein, J.** 2004. *Inferring Phylogenies*, First ed. Sinauer Associates, Inc.,  
353 Sunderland, Mass.
- 354 10. **Felsenstein, J.** 2005. PHYLIP (Phylogeny Inference Package) version 3.6. . Distributed by  
355 the author. Department of Genome Sciences, University of Washington, Seattle.
- 356 11. **Finn, R. D., J. Mistry, J. Tate, P. Coghill, A. Heger, J. E. Pollington, O. L. Gavin, P.**  
357 **Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. Sonnhammer, S. R. Eddy, and A.**  
358 **Bateman.** 2010. The Pfam protein families database. *Nucleic Acids Res.* **38**:D211-D222.
- 359 12. **Fouts, D. E., L. Brinkac, E. Beck, J. Inman, and G. Sutton.** 2012. PanOCT: automated  
360 clustering of orthologs using conserved gene neighborhood for pan-genomic analysis  
361 of bacterial strains and closely related species. *Nucleic Acids Res.* **40**:e172.
- 362 13. **Kaas, R. S., C. Friis, D. W. Ussery, and F. M. Aarestrup.** 2012. Estimating variation  
363 within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli*  
364 genomes. *BMC Genomics* **13**:577.
- 365 14. **Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori.** 2004. The KEGG  
366 resource for deciphering the genome. *Nucleic Acids Res.* **32**:D277-D280.
- 367 15. **Klimke, W., R. Agarwala, A. Badretin, S. Chetvernin, S. Ciufu, B. Fedorov, B. Kiryutin,**  
368 **K. O'Neill, W. Resch, S. Resenchuk, S. Schafer, I. Tolstoy, and T. Tatusova.** 2009. The  
369 National Center for Biotechnology Information's Protein Clusters Database. *Nucleic*  
370 *Acids Res.* **37**:D216-D223.
- 371 16. **Koonin, E. V.** 2005. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*  
372 **39**:309-338.
- 373 17. **Koonin, E. V., and Y. I. Wolf.** 2008. Genomics of bacteria and archaea: the emerging  
374 dynamic view of the prokaryotic world. *Nucleic Acids Res.* **36**:6688-6719.
- 375 18. **Koski, L. B., M. W. Gray, B. F. Lang, and G. Burger.** 2005. AutoFACT: an automatic  
376 functional annotation and classification tool. *BMC Bioinformatics* **6**:151.

19. **Kristensen, D. M., L. Kannan, M. K. Coleman, Y. I. Wolf, A. Sorokin, E. V. Koonin, and A. Mushegian.** 2010. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics* **26**:1481-1487.
20. **Kristensen, D. M., Y. I. Wolf, A. R. Mushegian, and E. V. Koonin.** 2011. Computational methods for Gene Orthology inference. *Brief. Bioinform.* **12**:379-391.
21. **Laing, C., C. Buchanan, E. N. Taboada, Y. Zhang, A. Kropinski, A. Villegas, J. E. Thomas, and V. P. Gannon.** 2010. Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinformatics* **11**:461.
22. **Lappin, E., and A. J. Ferguson.** 2009. Gram-positive toxic shock syndromes. *Lancet Infect. Dis.* **9**:281-290.
23. **Lefebure, T., V. P. Richards, P. Lang, P. Pavinski-Bitar, and M. J. Stanhope.** 2012. Gene repertoire evolution of *Streptococcus pyogenes* inferred from phylogenomic analysis with *Streptococcus canis* and *Streptococcus dysgalactiae*. *PLoS One* **7**:e37607.
24. **Lefebure, T., and M. J. Stanhope.** 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol.* **8**:R71.
25. **Li, L., C. J. Stoeckert, and D. S. Roos.** 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**:2178-2189.
26. **Marchler-Bauer, A., C. Zheng, F. Chitsaz, M. K. Derbyshire, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, C. J. Lanczycki, F. Lu, S. Lu, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, and S. H. Bryant.** 2013. CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.* **41**:D348-D352.
27. **Pagani, I., K. Liolios, J. Jansson, I. M. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides.** 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **40**:D571-D579.
28. **Rivera, M. C., R. Jain, J. E. Moore, and J. A. Lake.** 1998. Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. U. S. A.* **95**:6239-6244.
29. **Roth, A. C., G. H. Gonnet, and C. Dessimoz.** 2008. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9**:518.
30. **Schmitt, T., D. N. Messina, F. Schreiber, and E. L. Sonnhammer.** 2011. Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Brief. Bioinform.* **12**:485-488.
31. **Schneider, A., C. Dessimoz, and G. H. Gonnet.** 2007. OMA Browser--exploring orthologous relations across 352 complete genomes. *Bioinformatics* **23**:2180-2182.
32. **Snipen, L., T. Almoy, and D. W. Ussery.** 2009. Microbial comparative pan-genomics using binomial mixture models. *BMC Genomics* **10**:385.
33. **Snipen, L., and D. W. Ussery.** 2010. Standard operating procedure for computing pangenome trees. *Stand. Genomic Sci.* **2**:135-141.
34. **Sonnhammer, E. L., and E. V. Koonin.** 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet* **18**:619-20.
35. **Starr, C. R., and N. C. Engleberg.** 2006. Role of hyaluronidase in subcutaneous spread and growth of group A *Streptococcus*. *Infect. Immun.* **74**:40-48.
36. **Suzek, B. E., H. Huang, P. McGarvey, R. Mazumder, and C. H. Wu.** 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* **23**:1282-1288.
37. **Suzuki, H., T. Lefebure, M. J. Hubisz, P. Pavinski Bitar, P. Lang, A. Siepel, and M. J. Stanhope.** Comparative genomic analysis of the *Streptococcus dysgalactiae* species group: gene content, molecular adaptation, and promoter evolution. *Genome Biol. Evol.* **3**:168-185.

429 38. Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D.  
430 M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A.  
431 V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale. 2003. The COG  
432 database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**:41.

433 39. Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward, S. V.  
434 Angiuoli, J. Crabtree, A. L. Jones, A. S. Durkin, R. T. Deboy, T. M. Davidsen, M. Mora,  
435 M. Scarselli, I. Margarit y Ros, J. D. Peterson, C. R. Hauser, J. P. Sundaram, W. C.  
436 Nelson, R. Madupu, L. M. Brinkac, R. J. Dodson, M. J. Rosovitz, S. A. Sullivan, S. C.  
437 Daugherty, D. H. Haft, J. Selengut, M. L. Gwinn, L. Zhou, N. Zafar, H. Khouri, D.  
438 Radune, G. Dimitrov, K. Watkins, K. J. O'Connor, S. Smith, T. R. Utterback, O. White,  
439 C. E. Rubens, G. Grandi, L. C. Madoff, D. L. Kasper, J. L. Telford, M. R. Wessels, R.  
440 Rappuoli, and C. M. Fraser. 2005. Genome analysis of multiple pathogenic isolates of  
441 *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl.*  
442 *Acad. Sci. U. S. A.* **102**:13950-13955.

443 40. Tettelin, H., D. Riley, C. Cattuto, and D. Medini. 2008. Comparative genomics: the  
444 bacterial pan-genome. *Curr. Opin. Microbiol.* **11**:472-477.

445 41. Welch, R. A., V. Burland, G. Plunkett, 3rd, P. Redford, P. Roesch, D. Rasko, E. L.  
446 Buckles, S. R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou,  
447 D. C. Schwartz, N. T. Perna, H. L. Mobley, M. S. Donnenberg, and F. R. Blattner. 2002.  
448 Extensive mosaic structure revealed by the complete genome sequence of  
449 uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **99**:17020-17024.

450 42. Wescombe, P. A., M. Upton, K. P. Dierksen, N. L. Ragland, S. Sivabalan, R. E.  
451 Wirawan, M. A. Inglis, C. J. Moore, G. V. Walker, C. N. Chilcott, H. F. Jenkinson, and J.  
452 R. Tagg. 2006. Production of the lantibiotic salivaricin A and its variants by oral  
453 streptococci and use of a specific induction assay to detect their presence in human  
454 saliva. *Appl. Environ. Microbiol.* **72**:1459-1466.

455 43. Willenbrock, H., P. F. Hallin, T. M. Wassenaar, and D. W. Ussery. 2007.  
456 Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome  
457 microarray. *Genome Biol.* **8**:R267.

458 44. Wolf, Y. I., and E. V. Koonin. 2012. A tight link between orthologs and bidirectional  
459 best hits in bacterial and archaeal genomes. *Genome Biol. Evol.* **4**:1286-1294.

460 45. Zhao, Y., J. Wu, J. Yang, S. Sun, J. Xiao, and J. Yu. 2012. PGAP: pan-genomes analysis  
461 pipeline. *Bioinformatics* **28**:416-418.

462  
463  
464

465

## 466 LEGENDS TO FIGURES

467 **Figure 1.** GET\_HOMOLOGUES flowchart and its outcomes. BLAST and optional Pfam  
468 searches are optimized for local (multi-core) and cluster computer environments.  
469 While BDBH uses one sequence from the reference genome to grow clusters, COG  
470 requires a triangle of reciprocal hits. Instead, OMCL groups nodes in a BLAST graph to  
471 build clusters. Note that these clustering algorithms can be fine-tuned by customizing  
472 parameters such as -C (min. %coverage in pairwise BLAST alignments), -E (max. E-value  
473 for a hit to be considered), -D (require equal Pfam domain composition when defining  
474 similarity-based orthology composition), -S (min. %sequence identity in BLAST  
475 query/subject pairs [BDBH|OMCL]) and -N (min. BLAST neighborhood correlation  
476 [BDBH|OMCL]). In addition, the user can choose which genome should be used as the  
477 reference using option -r.

478

479 **Figure 2.** Pangenome analysis of 50 *Streptococcus* genomes from 14 species. A) Venn  
480 diagram of core-genomes generated by the BDBH, COG and OMCL strategies. B) Core-  
481 genome size estimate with the Tettelin (blue) and Willenbrock (red) fits (39, 43). C)  
482 Pan-genome size estimate with the Tettelin fit. D) Venn analysis of pan-genomes  
483 generated by COG and OMCL. E,F) Partition of OMCL pan-genomic matrix in shell,  
484 cloud, soft-core and core compartments. These plots can be easily created with  
485 GET\_HOMOLOGUES accompanying scripts, as explained in the manual.

486

488 **Figure 3.** Parsimony pangenome tree for 50 *Streptococcus* proteomes derived from  
489 presence/absence data in a consensus (OMCL & COGtriangles) pangenome matrix  
490 computed from the OMA50 dataset, as detailed in the main text. This phylogeny was  
491 the most parsimonious tree found in a tree search performed with PARS from the  
492 PHYILIP suite, using 50 data jumbles. The tree has a total length of 11473 steps.







